DOCUMENT RESUME

ED 266 684 FL 015 540

AUTHOR Raatz, Ulrich
TITLE Tests of Reduced Redundancy--The C-Test, A Practical Example.
PUB DATE 85
NOTE 15p.; In: Language Testing in School. AFinLA Yearbook 1985. No 41 (see FL 015 537). See also FL 015 539. Document is in small print.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cloze Procedure; Educational Research; Elementary Secondary Education; Foreign Countries; Language Proficiency; *Language Tests; *Redundancy; Second Language Learning; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS *Authentic Materials

ABSTRACT

Classical cloze tests have been criticized as unsatisfactory operationalizations of the concept of reduced redundancy and the theory of expectancy grammar. Use of the C-Principle is proposed as an improvement over the classical cloze procedure. Such tests are constructed according to a technique that deletes the second half of every second word in authentic texts, beginning with the second word in the second sentence and skipping single-letter words. Four to six similarly treated texts are combined in a C-Test. Only entirely correct restorations by the examinee are acceptable. The advantage of this technique is that redundancy is reduced at the word level rather than at the sentence level. The texts chosen should be brief, authentic, and neutral in subject matter. Tests for beginning learners are to be constructed from simpler texts, with a generous amount of time allowed for completion. Because of the authenticity of texts, item analysis can be performed only at the text level; it is defined as the percentage of correct solutions. In several years of testing in several languages (native, second, and foreign), the reliability and validity ratings of the technique have ranged from acceptable to high. Investigations have also established its construct validity. C-Tests can be used for student placement and diagnosis, and as research instruments. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

ED266684

TESTS OF REDUCED REDUNDANCY - THE C-TEST, A PRACTICAL EXAMPLE

Ulrich Raatz (University of Duisburg)

1. Criticism of cloze tests

Cloze tests are unsatisfactory operationalizations of the concept of reduced redundancy, and therefore also of the theory of expectancy grammar, as Alderson (1978) and Klein-Braley (1981) have shown. In addition to the criticism mentioned, they also pointed out:

1. The deletion rates used in classical cloze tests are too high. If 50 items are usually demanded in order to ensure an acceptable level of reliability then extremely long texts are necessary. In addition the deleted words are not a representative sample of different word classes of the text. This means that cloze tests developed from the same text but with different starting points and deletion rates differ in terms of difficulty, reliability and validity, as Klein-Braley showed quite clearly in her study.

2. If one uses only one text, as is standard practice in cloze tests, one cannot assume that this is a representative sample of the language.

3. The two different scoring methods for cloze tests are problematical. If the exact method is used cloze tests are too difficult even for competent native speakers. But the acceptable scoring method is much less objective and extremely time-consuming. This is true especially for the cloze-entropy scoring technique (Darnell 1970).

4. The basic requirement that adult educated native speakers of a language should score full marks on a tests intended for non-natives is not fulfilled by any cloze tests.

This criticism led to the establishment of criteria which an improved procedure ought to fulfill, although these criteria were pure theory and seemed utopian. They were:

1. The new test should be much shorter, but at the same time it should have at least 100 items.

2. The deletion rate and the starting point for deletions should be fixed, and no longer a matter of personal preference.

3. The words affected by the deletions should be a genuinely representative sample of the elements of the text.

FL015540

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Tomola

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

4. Examinees with special knowledge should not be favoured by specific texts, in other words the new test ought to contain a number of different texts.

5. Only exact scoring should be possible so as to ensure objectivity.

6. Native speakers ought to be able to score 100% on the test.

In addition the test, of course, nad do be reliable and valid and very easy to develop.

These requirements seem impossible to fulfil if the classical cloze principle is retained. The principle must be modified. One modification is the C-Principle. The tests constructed using this principle are a very much better operationalization of the underlying theory than classical cloze tests.

2. C-Principle and C-Tests

It is obviously the deletion technique which leads to the difficulties encountered with the classical cloze test. This was the reason that we decided to start here in our attempts to improve the tests. We developed a new deletion technique which we called the C-Principle. Deletions are not performed at the sentence level but at the word level. We no longer remove whole words from the text; we "damage" parts of words.

How does the C-Principle work? In short, carefully selected, preferably authentic texts we delete the second half of every second word, beginning with the second word of the second sentence. If the word has an odd number of letters then half a letter more is deleted. If the word only has one letter, as, for instance, in the English words "a" and "I", then this word is ignored in the counting. A number of texts which have been mistreated in this way, usually between 4 and 6, are put together to form a "C-Test". The examinee is confronted with these texts in the form of a test paper and is asked to restore the deleted elements. He is given a reasonable amount of time to perform this task. Only entirely correct restorations are counted as right.

The C-Tests which have been developed according to this principle have the following substantial advantages over cloze tests:

1. Many more items are possible with much shorter texts. A classical cloze test using a 5th word deletion rate would have to be at least 500 words long in order to contain 100 items. A C-Test consisting of 5 texts with 20 half deleted words would be only approximately half as long.

2. C-Test scoring is "exact" and objective because there is almost always only one possible solution. In a few exceptional cases there are a small number of different solutions

(e.g. this and that). In such cases all solutions are accepted.

3. The scoring of a C-Test is very quick and easy for a native speaker. He or she only needs to read through the text once, and normally needs no key.

4. C-Tests are very "easy" for native speakers, because the first half of the word is left standing. In a number of different pilot studies it has been shown that native speakers were normally able to restore over 95% of the damaged words. On the other hand, someone who does not understand the language normally scores zero on the test.

5. Since every second word is "damaged" the probability of obtaining a representative sample of all the word classes in the text is very much higher. This was examined in a special investigation,which was aimed at revealing whether deletion of every second word in a 200 word text produced a representative sample of all the words in the text. Two students were asked to examine 40 English and 40 German texts, each with 200 words, chosen at random to vary as much as possible so far as style and content were concerned. They coded each text into the categories: nouns, adjectives, verbs, adverbs, prepositions, articles and remainder. Then they deleted every second word and compared the distribution of the deleted elements to that of the elements in the original texts. The results showed that for every one of the 80 texts the deleted samples are representative for the entire text.

6. Because a C-Test consists of a number of different neutral texts the sampling of content classes is better. Test subjects who happen to have special knowledge in certain areas have no substantial advantage over the other subjects.

C-Tests are an operationalization of the theory of reduced redundancy. This one can see clearly if one looks at the overview on page . In contrast to the other procedures a C-Tests consists of 5 to 6 short texts. Deletion takes place at word level. The channel is visual, as in the cloze test, and the examinee behaviour is also similar to that desmanded by the cloze test, although here one has to use the term gaps rather than blanks to indicate that only parts of words have to be restored.

There is however one mayor difference. Redundancy is present in language on a number of different levels: on the level of santances, of words, and naturally on the level of the letters inside the word. This third type of redundancy can be used in the C-Test too, to help in finding the appropriate solution. The processes which take place inside the subject when he or she is working through a C-Test are currently under investigation.

Using the C-Principle, C-Tests have been developed in several different languages. A variety of subject groups have been tested, and a number of different questions have been investigated. The languages belong typologically to the analytic and synthetic classes: German, English, French, Spain, Turkish, Dutch and Ivrit. However no studies have been conducted so far with a genuinely agglutinating language such as Finnish.

The following figures show examples of parts of C-Tests in two different languages:

Figure 1

Part of a German C-Test

Gestern ist eine 76 Jahre alte Frau in Garath bei einem Unfall todlich verletzt worden. Die Fußga_ _ _. hatte i_____ Höhe d_ _ . Stettiner Str_ _ _ die Frank_ _ _. Straße überq_ wollen u_ _ _ _ war da_____ von ei_ _ Auto angef_____ worden. D_____ Ampelanlage a dieser Straß _ _ war z_____ Zeitpunkt d_ _ _ Unfalls abges_ _ _ und wu_____ repariert. I_ _ _ Krankenhaus er_____ die Frau ihren schweren Kopfverletzungen.

Figure 2

Part of an English C-Test

Pollution is one of the big problems in the world today. Towns a_ cities a _ growing, indu_ is gro . and t population o_ _ the wo_ is gro Almost every_ causes poll _ in so way o_ _ another. T _ _ air i filled wi fumes fr _ factories a vehicles, a_ _ there i _ noise fr aeroplanes a_ _ machines. Riv . lakes a seas a_ _ polluted b factories and by sewage from our homes.

BEST COPY AVAILABLE

In the English versions each text has 25 deletions, because English words tend to be shorter on average. This phenomenon had been adjusted for by raising the number of deletions.

3. How is a C-Test developed?

First of all one looks for a number of suitable short texts in the appropriate language which are complete in themselves. The texts should preferably be authentic, and neutral in their subject matter. In developing tests for schoolchildren in their native language we have found school text books very useful, but of course these must be ones which are not known to the target group. Readers in particular are a rich source for suitable texts. Texts for students or adult groups were obtained from reference books and other nonfiction sources, or newspapers.

One then deletes 20 or 25 half words from the texts. The deletions normally begin in the second word of the second sentence. At the end the text runs on until it reaches a natural conclusion.

Four to six texts which have been treated like this are assembled to form a C-Test. Just as in the construction of any test it is a good idea to order the texts in ascending difficulty.

The time allowed for working through the test should be generous. We have found that in general 5 to 7 minutes are needed for each text in the test.

To score the text the scorer reads through the text and marks the correct solutions. The total score on the test is the sum of the part scores. In those cases where items have more than one possible solution all these solutions are counted as correct. Our experience has shown that this method of scoring is objective and fairly quick to carry out.

Problems can arise if C-Tests are to be developed for foreign language learners in the early stages of learning. Subject groups of this kind have only a very limited vocabulary, and this is defined by the textbook they are using. The same applies to their repertoire of structures. If C-Tests for such groups were constructed from authentic texts they would be both too difficult and unfair. In such cases we recommend the analysis of the vocabulary and structures used in the textbook, followed by the construction of quasi-authentic texts by native speakers, using only those words and structures which are known to the pupils. The other alternative is to use texts taken from textbook series which are not being used by the target group, but which use the same linguistic material. Here, of course, modifications can be necessary.

Item analysis in the classical sense, i.e. on the deletion level, which usually takes place after the first trial run and which is used to improve the test, is neither possible nor desirable with the C-Test. This is because the texts used are authentic, and it would not be possible to delete unsuitable items without

BEST COPY AVAILABLE

destroying both the sense of the text and the deletion principle. Furthermore the items cannot be statistically independent of each other, since they are all embedded in the same linguistic context. The only way a C-Test can be improved is by deleting or changing an entire text.

This means that item analysis for C-Test must be performed on the text or super-item level. In this analysis the difficulty of a text (superitem) can be defined as the percentage of correct solutions, the discrimination as the part-residual correlation of the super-item with the total score, and the reliability as Alpha-reliability according to Cronbach. In order to examine the question of whether all the texts in a C-Test measure the same overall proficiency factor analysis is not a suitable technique (Raatz, 1982). A better method is the use of the CLA model developed by Moosbrugger and Müller (1982) which is similar to the well-known Rasch model.

4. Investigations into the reliability and validity of C-Tests

When a new approach to the measurement of general language proficiency is proposed, an approach which claims to be suitable for a variety of languages a number of different target groups including L1, L2 and LF learners, and to be useful for a variety of purposes, then it is essential to perform a large number of empirical investigations, varying all possible conditions. The languages, the groups, the age levels and the problems must all be investigated to discover whether the C-Principle really produces reliable, valid and useful tests. If it can be shown that one particular C-Test, for instance a German C-Test, developed for a specific target group and to be used for a specific prupose, is sufficiently reliable and valid, this is very gratifying for the researchers, but it says nothing about the generalisability of the C-Principle.

For this reason we have spent the last few years performing a large number of investigations into the performance of C-Tests for a variety of languages, groups, age levels and problems. Other colleagues, both inside and outside of Germany, have performed their own investigations.

In virtually all cases reliability was estimated using Cronbach's Alpha. It is common knowledge that this formula underestimates the measurement accuracy of a test procedure.

In order to answer the question of what a C-Test actually measures, concurrent and predictive validity were investigated, in some cases factorial and construct validity were examined. Sometimes our external criteria were school grades in the appropriate language, sometimes they were ratings of the general level of proficiency attained given by teachers, in some cases we had the results of other test batteries which had been administered at the same time. In the investigations of factorial validity we used among other things the results of other language tests and intelligence tests.

4.1. C-Tests with L1 groups

Table 1 shows the results of 6 major investigations into reliability and validity with L1 groups. These studies were done by us with the help of students at our university and local teachers. The language of the C-Test was German in the first four investigations, English in the fifth and Turkish in the sixth. The pupils came from the second to eleventh grades. The third column shows the number of subjects involved in each investigation, the fourth the reliability coefficients. In those cases where a number of coefficients were obtained the lowest and highest are given, or the median of all reliability coefficients. The fifth column shows the concurrent validity. The criteria are school grades and teacher ratings. In one case we used pupil self assessments. In the second investigation the criterion was determined six month after the test administration.

Table 1

Reliability and validity of C-Tests
for different groups and languages

L1 Groups

Language	Sample	N	r_{tt}	r_{tc}
German	schoolchildren 3rd - 8th grade	716	.76 - .88	.47 - .80 school grades
German	schoolchildren 2nd grade	117	.83	.63 (med) ratings (half year later)
German	schoolchildren 5th grade	75	.78 - .81	.57 - 65 school grades factorial
German	schoolchildren 6th - 11th grade	402	.78 (med)	.40 (med) school grades
English	schoolchildren 10 - 15 years	160	.66 - .92	.73 - .88 ratings/self assessment
Turkish	schoolchildren 17 years	13	.90	.45 ratings

At the last two columns show, the reliability coefficients range between acceptable and high, the validity coefficients are most satisfactory.

4.2. C-Tests with L2 groups

Table 2 shows the results of six investigations into general language proficiency in German as a second language (5 investigations) and Hebrew as a second language. Studies 1 to 5 were in the main conducted by Edgar Süßmilch from Düsseldorf in cooperation with us. The investigations into adult learners of Hebrew as a second language were performed by Andrew Cohen in Israel. The children involved in studies 1 to 3 are the children of immigrant workers living in Germany. Again the third column shows the size of the sample involved, the fourth the reliability coefficients, some of which are remarkably high, and the fifth column shows an estimate

Table 2

Reliability and validity of C-Tests
for different groups and languages

		L2 Groups		
Language	Sample	N	r_{tt}	r_{tc}
German	Turkish children 2nd - 4th grade	227	.85 - .94	.41 - .88 school grades and ratings
German	Greek children 3rd - 4th grade	305	.82 - .92	.58 - .90 school grades and ratings
German	other nations	87	.83 - .88	.73 school grades and ratings
German	Turkish pupils 17 years	15	.86	.72 school grades
German	Turkish girls	37	.76	.57 ano test (3 subtests)
Hebrew	adults	62	.93 - .95	.69 - .87 (lang. tests) .72 - .92 (ratings)

of concurrent validity. Again the criteria are ratings and school grades with the exception of the last two studies where other language tests were used. Here again the results are very positive.

4.3. C-Tests with LF groups

Table 3 shows the results of 7 investigations which we either conducted ourselves or which were reported to us.

The first investigation shows the results obtained for a multinational group of young people and adults learning German at a language school. The criterion for this investigations was the EUROTEST, a discrete-point item test with 4 subtests used for placement in the language school. Reliability is very high, and so is the agreement between the C-Test and the discrete-point item test.

Table 3

Reliability and validity of C-Tests
for different groups and languages

		FL Groups		
Language	Sample	N	r_{tt}	r_{tc}
German	Foreign students	252	.92 - .93	.82 - .86 Eurotest (4 Subtests)
English	German school-children 6th grade	108	.81	.60 school grades
English	German students	438	.77 - .91	.62 - .90 Delta Test (6 Subt)
English	Finnish students	51	.87	
English	Belgian students	192	.77	
French	German students	79	.97	.87 Bochum French Plac. Test (7 Subt)
Spanish	German adults (Further educ)	24	.92 - .94	.90 - .95 ratings

The second investigation shows the results obtained for 108 German pupils learning English in the second year of secondary school. For this study a C-Test consisting of pseudo-authentic texts using only the vocabulary and structures known to the pupils and written by native speakers of English was developed.

The subjects in the third study are students of English at the University of Duisburg who have to take a placement test before being assigned to the appropriate language courses. They took a C-Test together with the DELTA test, a discrete-point item test with 6 subtests.

The next two investigations were performed with Finnish students by Hilkka von Essen and with Belgian students by Bea von Beckhoven. We did the computing. For these investigations we have no validity coefficients.

The last investigation was performed by Grotjahn and Stemmer in Bochum. The subjects were 79 beginners in the French department taking the Bochum placement test in French, again a discrete-point item test, which was administered together with the C-Test.

The Spanish test, which we developed ourselves, was used in a pilot study. Here the subjects were adults learning Spanish at the local evening institute in their spare time.

As Table 3 shows, the results of the investigations with LF groups reveal perhaps even more clearly than the ones carried out with L1 and L2 groups that C-Tests in different languages and for different target groups are reliable, at times very highly reliable tests, and that in all cases they have high concurrent or predictive validity. We find it particularly interesting that C-Tests used for placement purposes correlate so highly with the results obtained with discrete-point item tests. This could lead to interesting developments in placement procedures.

4.4. Investigations into construct validity

In our recent work with the C-Test we have been examining various aspects of construct validation. The C-Test is based on the same fundamental theory as cloze tests. We posit an internalized grammar, or rule system, which is in the course of evolving in the learner. At any given point of measurement this system is self-consistent. As the learner progresses, he or she revises and adds to the system so that it becomes more and more similar to that of the adult educated native speaker of a language. By using a test of reduced redundancy in an acceptable operationalization we can force the learner to activate the internalized grammar in order to restore the missing elements in the text. The higher the score on the test, the more accurate, we conclude, is the learner's grammer, i.e. we draw conclusions from the learner's performance about his or her competence.

We tried to set up a number of predictions which followed from this theory and to examine whether they could be empirically confirmed.

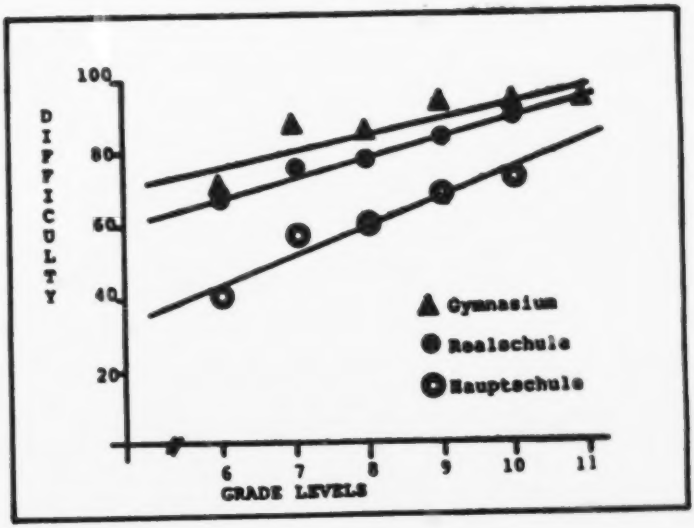
Our first hypothesis involved the basic question of whether C-Test performance was similar to performance in other language tasks. This, of course, is simple concurrent validation, and the results were positive.

We then turned our attention to age. For L1 groups, we thought, test scores ought to improve with increasing age. We could show, that relationship turned out to be linear (see Fig. 3).

The same study showed that the mean performance of pupils in the three different types of secondary schools in Germany differed, and that the results were in the direction predicted.

Figure 3

Empirical text difficulty levels and regression on age for pupils at different types of German secondary schools



Our next question was the rôle played by intelligence. We expected a moderate positive relationship between IQ scores and C-Test performance. We were able to show that this was the case, and also that C-Test scores in a factorial study clustered with other language scores (grades and a diagnostic test in L1, grades in LF) which indicated that a general non-language-specific verbal ability could be involved. The second factor was an intelligence factor with a speed component. (Raatz, 1984)

Klein-Braley investigated the hypothesis that, in terms of C-Test processing, texts had an inherent constant level of difficulty so that the observed C-Test score for any individual was a function of text difficulty and subject ability. It ought therefore to be possible to discover text characteristics which would enable us to predict test difficulty for specific groups. Again the results were positive. (Klein-Braley, 1984)

I was then able to show by use of a special latent trait model that this function for any given subject was an additive one. (Raatz, 1985)

We also investigated the question of the transferability of the C-Principle to other languages. We began our investigations with English and German. Grotjahn added French. I could show that the principle worked for Spanish. These are all Indo-European languages, and so we wondered whether it might be some particular feature of this language family which was responsible for the results. We felt that the validity of the principle received its confirmation when we were informed by Cohen (1985) that a satisfactory test had been constructed in Hebrew, and by Süßmilch's development of a test in Turkish.

So although the construct validation of C-Tests is still continuing - we are particularly interested in some work going on in Bochum (Stemmer, 1984) where C-Test processing strategies are under investigation - we feel at this point that we can claim construct validity for the C-Test.

5. Where can C-Tests be useful?

According to the underlying theory C-Tests provide reliable and valid information about the general level of proficiency in the native language or in a foreign language; they enable us to draw conclusions about the state of development an individual's competence in a language has reached from the assessment of his or her performance. However they provide no detailed information about strengths and weaknesses in individual areas of the language. So they tell us nothing about specific features of grammar, or syntax, or vocabulary, or spelling, and they offer no direct evidence about performance in listening and speaking. For this reasons the possible uses of C-Tests are limited.

For schools, universities and institutes of adult and further education we see the following four main areas where C-Tests could provide useful information:

1. C-Tests can be used at the beginning of a school year, a university term or a teaching unit as placement tests for determining which of a number of possible language courses is most suitable for an individual learner. They could also be used to select those students with very high proficiency levels in foreign languages.

2. It has proved difficult to develop suitable test procedures for the children of immigrant workers learning a second language. Such tests are needed do decide whether children should be given special tuition in the second language or whether they are able to attend normal classes together with L1 speakers of the language. Similar problems can also arise for adults, for instance in Isreal.

3. C-Tests can be used as global diagnostic instruments of learning difficulties in language, both in the native language (dyslexia) and in the foreign language. They would, of course, need to be backed up by other diagnostic procedures.

4. Since the same C-Tests can be used with L1, L2 and Fl learners they can be used in comparative studies as research instruments, for instance to investigate questions of language developement and language learning.

6. Conclusions

It appears that the C-Test does everything that the cloze test promised. Speaking as the researchers who conducted the vast majority of the investigations reported in this paper we can state that we are satisfied with the performance of the tests for all the languages and for all the groups with which they have been used. Nevertheless we must stress that a C-Test is not guaranteed to be automatically reliable and valid. It seems probable that, provided the C-Test is suitable in its level of difficulty for the target group envisaged, it will, in the majority of cases, produce valid results.

At the same time, before C-Tests are used in any institution for real decisions which would affect pupils' or students' lives, a certain amount of statistical evaluation is essential.

What must not happen is that uncricital users begin to abuse the C-Principle and the C-Tests by failing to perform the necessary statistical checking.

Bibliography

Alderson, J.C. 1978. The effect of certain methodological variables on cloze test performance and its implication for the use of the cloze procedure in EFL testing. Paper presented at the 5th International Congress of Applied Linguistics, Montreal, Canada.

Cohen, A.D., Segal, M. and Weiss, R. 1985. "The C-Test in Hebrew". In Klein-Braley, C. and Raatz, U. (eds). 1985. C-Tests in der Praxis. Bochum: AKS.

Darnell, D.K. 1970. Clozentropy: a procedure for testing English language proficiency of foreign students. Speech Monographs 37, 1: 36-46.

Grotjahn, J. and Stemmer, B. 19 . "On the development and evaluation of a C-Test for French". In Klein-Braley, C. and Raatz, U. (eds). 1985. C-Tests in der Praxis. Bochum: AKS.

Klein-Braley, C. 1981. Empirical investigation of cloze tests. An examination of the validity of cloze tests as tests of general language proficiency for German university students. Universität Duisburg: Inauguraldissertation.

- 1984. "Advance prediction of difficulty with C-Tests". In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (eds). 1984. Practice and problems in language testing 7. Colchester: University of Essex.

Moosbrugger, H. and Müller, H. 1982. "A Classical Latent Additive Test Model (CLA Model)". The German Journal of Psychology 6: 145-149.

Raatz, U. 1982. "Language theory and factor analysis". In Lutjeharms, M. and Culhane, T. 1982. Practice and problems in language testing 3. Proceedings of the Third International Language Testing Symposium of the IUS. Brüssel: Vrije Universiteit.

- "The factorial validity of C-Tests". In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (eds). 1984. Practice and problems in language testing 7. Colchester: University of Essex.

- "Better theory for better tests". To appear in Language Testing 1985.

BEST COPY AVAILABLE